CERN data services for LHC computing

To cite this article: X Espinal *et al* 2017 *J. Phys.: Conf. Ser.* **898** 062028

CERN data services for LHC computing

X Espinal, E Bocchi, B Chan, A Fiorot, J Iven, G Lo Presti, J Lopez, H Gonzalez, M Lamanna, L Mascetti, J Moscicki, A Pace, A Peters, S Ponce, H Rousseau and D van der Ster

CERN European Laboratory for Particle Physics 1211 Genève (Switzerland)

Abstract. Dependability, resilience, adaptability and efficiency. Growing requirements require tailoring storage services and novel solutions. Unprecedented volumes of data coming from the broad number of experiments at CERN need to be quickly available in a highly scalable way for large-scale processing and data distribution while in parallel they are routed to tape for long-term archival. These activities are critical for the success of HEP experiments. Nowadays we operate at high incoming throughput (14GB/s during 2015 LHC Pb-Pb run and 11PB in July 2016) and with concurrent complex production work-loads. In parallel our systems provide the platform for the continuous user and experiment driven work-loads for large-scale data analysis, including end-user access and sharing. The storage services at CERN cover the needs of our community: EOS and CASTOR as a large-scale storage; CERNBox for end-user access and sharing; Ceph as data back-end for the CERN OpenStack infrastructure, NFS services and S3 functionality; AFS for legacy distributed-file-system services. In this paper we will summarise the experience in supporting LHC experiments and the transition of our infrastructure from static monolithic systems to flexible components providing a more coherent environment with pluggable protocols, tuneable QoS, sharing capabilities and fine grained ACLs management while continuing to guarantee dependable and robust services.

1. Introduction

The data Storage and services group at CERN provides many services to the physics community. Here we describe two: CASTOR and EOS which are essential for LHC data taking, data processing, data distribution and analysis.

2. Storage systems for a broad scientific community: Castor and EOS

Storage system solutions for the broad scientific community is a challenging task. The four main LHC experiments are intensively using our two main systems: CASTOR [1] and EOS [2]. CASTOR is the Hierarchical Storage Management System for handling disk and tape layers. Born in 1999, it holds 190PB of data and 500M files in a common namespace. The current deployment status consists of six production instances: one for each major LHC experiment, one for the rest of the users and one dedicated to tape repacking activities. CASTOR fully relies on an ORACLE database backend which runs the namespace and handles the head nodes, disk servers and tape server logics. CASTOR's main role consists of data recording activities and data export and import. CASTOR's native protocol is xroot with gridftp support and phasing out legacy native protocol rfio. The current disk layout configuration is a mixture of software raid systems (mainly RAID-6 and RAID-60) and a brand new disk pool based on CEPH [4]. Access is secured via kerberos and X509 certificates.

During the First LHC Long Shutdown CASTOR evolved to central data recording functionalities and moved towards a colder storage system in front of the tape system. Some obsolete features were removed (ie. file updates, legacy protocols) and code simplified and optimised and part of the logics moved from C++ code to database PL/SQL procedures in the Database to run closer to the data. latest development on CASTOR were concentrated in improving per stream throughput performance to benefit from the available tape drive speeds ($>$ 350MB/s) and we developed xroot and gridftp plugins to be able to run CASTOR over a standard CEPH cluster via the RBD interface. This setup is now in production for targeted cases.

EOS started its production phase in 2011 and currently holds 158PB of data and 1.1B files. It is a disk-only storage solution mainly focused on analysis and fast data processing with a very low access latency thanks to the multi-replication across nodes and JBOD layout[1]. Fast metadata access is guaranteed by the in-memory resident per instance namespace[2]. XROOT [3] is the principal access protocol. Protocols such as gridftp, fuse mount and http are also supported. Authentication is done by Kerberos/X509. The four big LHC experiments (ALICE, ATLAS, CMS and LHCb) are using intensively EOS and they do have a *private* instance while for non-LHC experiments they make use of a *shared* instance, the main users (among others): AMS, COMPASS, NA48/62, nTOF, NA61 and CLIC. EOS is also the backend for the CERN file share and sync service: CERNBOX [8]

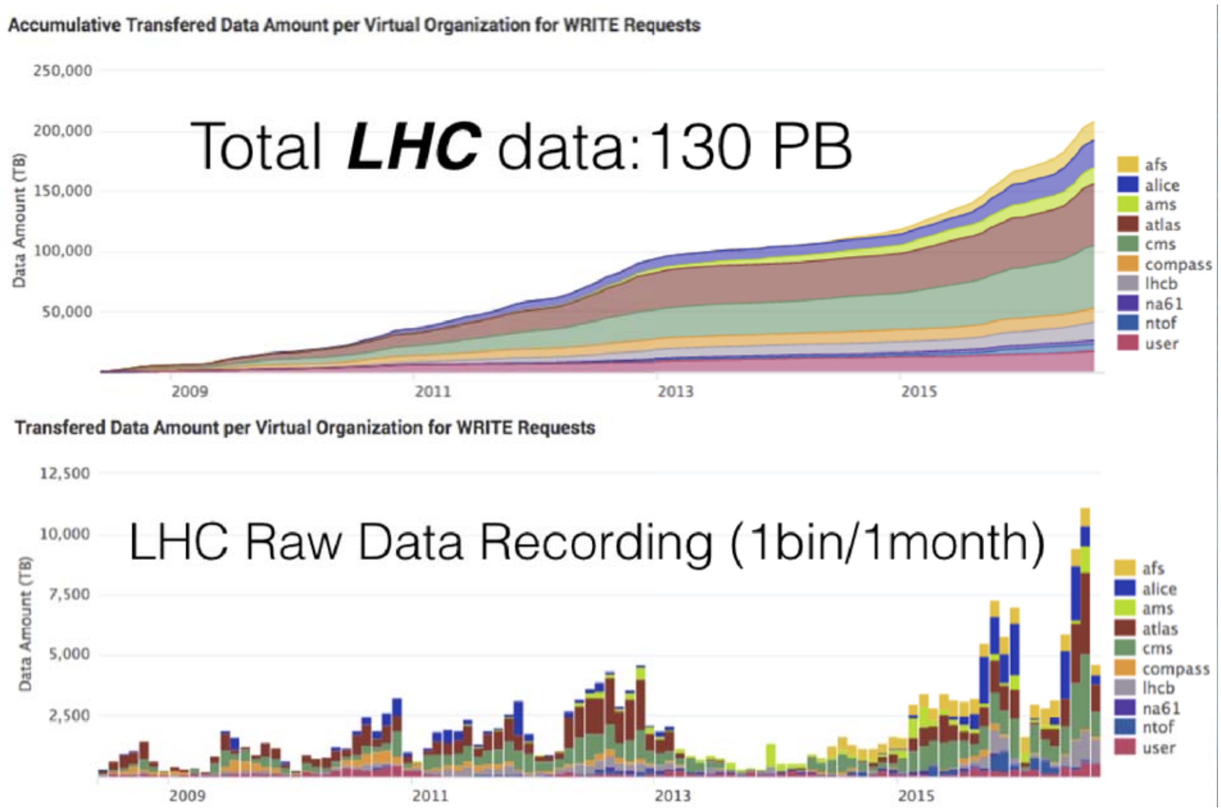

Figure 1. Data volume stored on tape since LHC start in 2008 until September 2016. The total volume of data is 190PB and LHC-only data accounts for 130PB. A data throughput record was achieved in July 2015 before the iCHEP conference with 10PB stored in four weeks

[1] The acronym stands for Just a Bunch Of Disks and can be seen as a collection of hard disks in a common enclosure that have not been configured to act as a redundant array of independent disks. The popularity of this layout has grown dramatically as raid systems performance are struggling with current commodity hard drive sizes [6]

[2] Currently evolving towards a distributed namespace infrastructure based on a shared K-V store system [7]

The data workflows are very similar at first order and can be summarised in different steps: detectors send data to our storage systems, then data is available for data processing and user analysis, and in parallel these data is replicated to other WLCG sites. One figure that can be taken out of this workflow is that for every byte this is coming to our storage systems from the detectors 4 bytes are *exported* over the network at the very beginning meaning that fresh data is very hot in the beginning before gradually cool down over time. Periodically not-so-new data need to be re-staged from the CASTOR tape system to be re-processed in order to apply new detector calibrations and geometries. This typically happens once or twice per year per experiment and the data volume to be recalled usually comprises all the RAW data taken during the run[3].

The LHC experiments are making a different use of CASTOR and EOS. On one hand ATLAS and CMS experiments prefers to stream data directly from the detector to EOS and then move it asynchronously to CASTOR for final tape recording using their data management frameworks (PanDA [5] and PhEDEX [9]) interleaved with the WLCG File Transfer Service FTS [10]. And on the other hand ALICE and LHCb preserved the operation mode from previous LHC run and data is sent from the detectors to CASTOR for tape recording and then moved to EOS for reprocessing, analysis and exportation. The different operation modes of the LHC experiments regarding data taking and data aggregation is clearly visible observing the throughputs in our systems (Figure 2).

Experiments have high-bandwidth links from the pits to IT computing centre and the data flow is critical to ensure coherent data recording. For this reason there is a disk buffer at the detector site to cover eventual issues with the data flow, nevertheless this is finite and problems in the transfer channel should be solved in due time to avoid losing events.

As previously mentioned ATLAS and CMS store data directly from the detector DAQ systems[4] to EOS but they do this in different ways. ATLAS is aggregating first the output in an online farm close to the DAQ infrastructure and then sending these data by bunches with peaks reaching 14GB/s whereas CMS is flowing constantly data with an average of 2GB/s during LHC operation (although peaks over 5GB/s has been observed during the proton-ion run).

ALICE and LHCB stream data to CASTOR with the particularity that ALICE requires high bandwidth during their data taking periods and more importantly during the heavy ion runs, they substantially fill the link from ALICE detector to IT, which consists of 8x10Gbps channels, with a data throughput between 60 to 70GB/s. In 2015 ALICE reached a systematic recording of 1PB per week on CASTOR during the Heavy Ion run which normally spans over three weeks. ALICE on the other hand make extensive use of EOS for world-wide data processing with O(10k) clients connecting from outside CERN driving the dedicated EOS disk pool to run between 10GB/s to 40GB/s depending on the analysis needs. Figure 2 shows the raw data recording at CERN by LHC, fixed target experiments and AMS. LHC data recorded so far accounts for 130PB.

2.1. Hardware and resources

The actual model to acquire new hardware is driven by the market and we are making extensive use of commodity hardware. The same type of diskservers are bought to serve all storage systems and they usually consist of a headnode with attached storage (the trend is 2 trays of 48 disks each). Memory and CPU are usually scaled to fit WLCG requirements of 2GB/core and networking moved towards the 10Gbps structure as the data volume stored in a single server is in the order of 300TB nowadays. Future deliveries can easily be made of 8TB to 10TB disks and 60 disks per enclosure and this is getting close to the figure of half a petabyte per unit. Table 1 summarise installed capacities and general usage measurements.

Besides the main LHC experiments we do provide the storage for smaller communities whose numbers used to be far from the big experiments but the paradigm is rapidly shifting, new detectors and newer DAQ systems can easily run at the scale of the large LHC experiments and the solutions this communities

[3] Last year 50PB of RAW data were recorded. Usually Tier-1 sites within the WLCG collaboration (Worldwide LHC computing Grid) take responsibility in reprocessing campaigns hence offloading CERN tape system

[4] Data Acquisition Systems

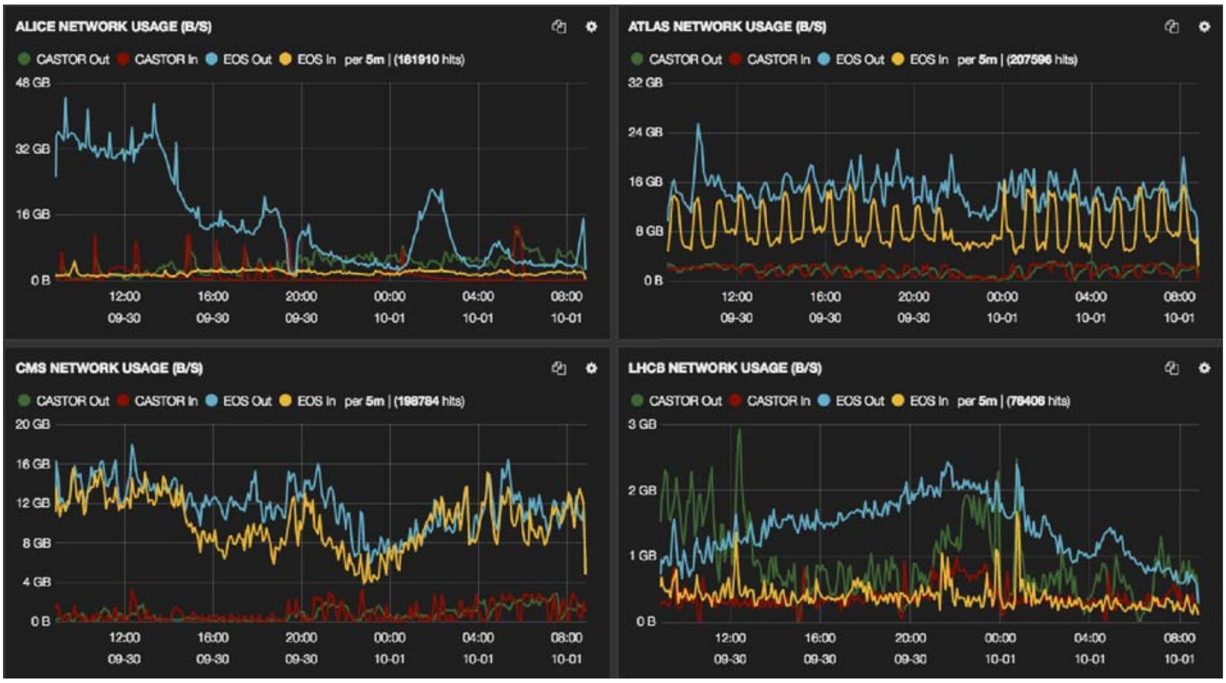

Figure 2. Different use of the Storage Systems by the LHC experiments. ALICE (top-left) and LHCB (bottom-right) stream data from the DAQ systems to CASTOR while ATLAS (top-right) and CMS (bottom-left) stream data directly to EOS

Table 1. CERN IT-ST resources and usage. Resources account for the installed capacities as per February 2017[†]*Global EOS numbers includes CERNBOX*[††]*CEPH work at object level the count for the number of files should be read as number of objects.*[‡] *CASTOR global numbers includes disk and tape, the numbers in brackets corresponds to the CASTOR disk cache.*

Service	Space(PB)	#files	#disks	iops		io$_{(peak)}$	clients$_{(avg)}$
EOS[†]	158	1.1B	50k	15k	write	20 GB/s	2k
					read	80 GB/s	50k
CEPH	12	0.5B[††]	5k	25k	write	5 GB/s	20k
					read	6.5 GB/s	
CASTOR[‡]	200(30)	0.5B(10%)	6k	-	write	12 GB/s	O(1k)
					read	24 GB/s	
CERNBOX	3	178M	50k	15k	write	200 MB/s	200
					read	150 MB/s	200

needs are not anymore different so they are using CASTOR and EOS very much in the same way as the biggest experiment do.

One of the main achievements during this year is the ability to mount the big data repositories on EOS on the batch nodes and interactive services. The users can browse the EOS namespace of their experiment and interact with the data on a close-to-file-system fashion, once they *log* into the node (batch or interactive) and try to access the data repository ie. $/eos/atlas$ the namespace is visible via an auto-mount and the user credentials are used to visualise the namespace tree where the user has access. The FUSE behaviour over EOS has been dramatically improved and is still a work in progress.

2.2. The importance of the network

Data taking on a multi-site infrastructure with impaired latencies between sites might lead to stream degradation in case there are losses at the network level. Single stream performance is severely impacted by packet drops on WAN (Wide Area Networks) with substantial latencies (CERN to WIGNER latency is 23ms). Stream throughput performance degrades as the TCP window falls to zero when packet drops are detected hence needing time to ramp up again to recover nominal link speeds. This affects some of the streams that are on the fly and could cause pile-up of slow transfers that eventually could block a percentage of the overall number of streams which is usually fixed. One reason we found among others is the saturation of the existing double link from CERN to WIGNER, the link consisting of two 100Gbps fibres[5] but full saturation has been observed periodically on the traffic from CERN to WIGNER Figure 3. Also accidents happen and a broken link drives to saturation and latencies started to dominate transfers once the capacity is reduced by half.

The actual deployment model of EOS is following a complete split approach between both sites in terms of hardware and replicas, so one replica is stored at CERN and a second replica is stored at WIGNER. The client receives the exit callback once the second copy is consolidated. This second copy is usually at Wigner as we favour locality of the clients for the first transfer so experiments DAQ systems always write first to Meyrin as this also favours the network optimisation as the client only talks from the detector pit to the CC in Meyrin.

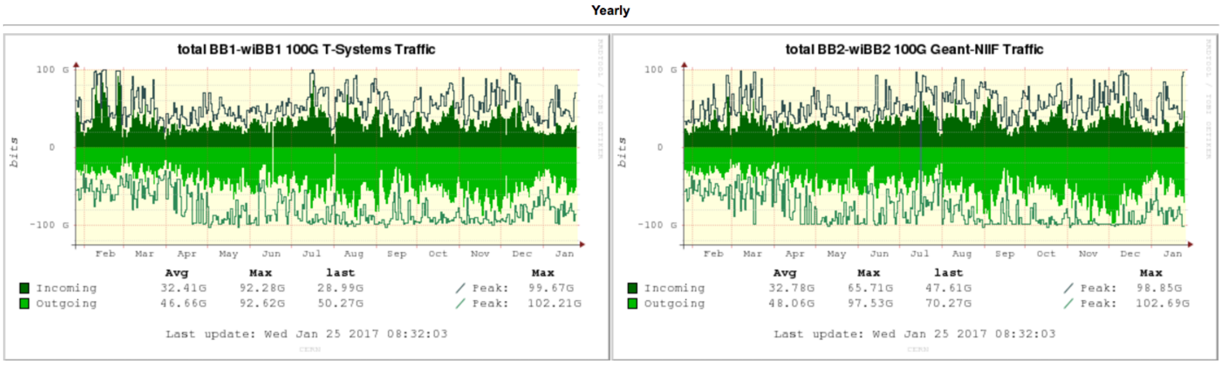

Figure 3. Meyrin-Wigner link occupancy reached its limits during the first phase of the Run-II. Packet drops on WAN with substantial latencies (23ms) impacts heavily on the stream throughput performance as the TCP window scaling drops to zero before scaling up again.

3. From the pit to the users: a complex machinery. Real production example of slowdown impact

On July ATLAS and CMS experiments reported a growing backlog of data to be exported at the Tier-1s was accumulating since some weeks. Experts from the different systems involved gathered together to evaluate the situation as the system was running but not at the required speed. What was found embraces: **Network**: spotted saturations (peaks) at the gridftp door level, CMS experiment were formally using 4x10Gbps gridftp doors and two more nodes were installed immediately. There was an immediate increase of the throughput rate peaking at the total design value of 6.5GB/s (maximum of almost 1.2GB/s

[5] At the time this article is being written a third link has been commissioned and speed already reached 300GB/s between CERN and Wigner data centers

per node) but not maintained. During the debugging session with network experts it was revealed a high number of TCP retransmits/discards on the Top of the Rack switches where gridftp doors (for all experiments) are located. This have a very bad impact for transfers involving Wigner nodes due to higher RTT (22ms). This is a known issue with the switch model and network group propose to install newer switch models that should offer line-rate forwarding on all ports. Following this some days later new switches were deployed and 8 gridftp doors were re-cabled. This had an immediate positive impact: neutralised the blocking factor and alleviate the effect of misbehaving switches when running at port capacity of around 60% (where we observed large quantity of drops/retransmits). The experiments had at this point 8 gridftp doors each which should translate into a throughput capacity of 8GB/s exporting data from CERN to the WLCG.
Gridftp doors: the transfer quality plots of CMS were showing a constant failure rates between 20% to 40% depending on the sites, on this issue we have found an old cronjob to kill stale ftp transfers installed on the gridftp doors was killing transfer after a 60 mins timeout, this was set around 2011 where networks links and speeds were quite different from what we have now. This was changed on the gridftp doors. Transfer quality rates improved.
FTS: transfers taking longer than 1 hour from EOS to dCache sites were canceled by dCache after 1 hour. The source of the problem was on FTS which was not filling a transfer request timeout on the SRM put requests. Transfer quality improved much more and get back to normal values.
Phedex: was filling FTS slots with workflows that were not using the channels efficiently. One of the sites that had a major backlog to get data from CERN was RAL so the the number of connection for the FTS channel CERN-RAL was increased and also CMS did tune the workflows to optimise it.
This situation demonstrate the complexity of the systems we are running and the need of communication and cooperation within the involved parties. Thanks to the fast reaction the issue was solved in approximately one week (Figure 4).

4. Storage Services are not only for physics: the CEPH deployment at CERN

The virtual machine provisioning service at CERN OpenStack had the needs of an underlying storage for the virtual images and for extended storage volumes (GLANCE and CINDER) this requirement was fulfilled by implementing a block storage service which we cover with CEPH. Right now there are five CEPH clusters which cover a variety of use-cases that are outside the scope of our two main Large Scale Storage Systems for Physics. The CEPH implementation at CERN is composed by the following production clusters: **Production cluster for Openstack** which is now 5.5 PB providing Cinder (block storage volumes) and Glance (images repo), Rados GW (object storage interface for S3, Swift), NFS services for CVMFS and other appliances that need high IOPS number (Puppet master nodes require to sustain peaks of 40kHz).
Production cluster for CASTOR backend consisting of 4.2 PB to provide disk buffer/cache in front of tape drives for faster streaming and the ability to exercise CEPH native access from the tape drives.
Storage platform for HPC cluster: they have strong POSIX needs and close to 100% uptime as they have O(month) wall clock jobs. Currently we are evaluating CephFS for a shared scratch space on 50 HPC nodes and has been running stable over six months on a cluster of 0.5PB and three replica schema.
ATLAS S3 Event Service consisting of an S3 interface consisting of 5x10Gbit S3 gateways to CEPH storage backend and 0.5PB (used so far: 80TB and 100M objects)

5. New paradigms in data access, processing and sharing: Cernbox

CERNBOX is a sync and share layer on top of EOS based on OwnCloud software suite and it provides a cloud synchronisation service for all CERN users between personal devices and centrally-managed data storage. The big advantage is that the personal area is mounted on the CERN interactive services (lxplus) and batch nodes hence opening the window of a new paradigm in the way the users access the data and the way to handle the analysis results. Analysis and collaborative work can be done on-the-fly as the user

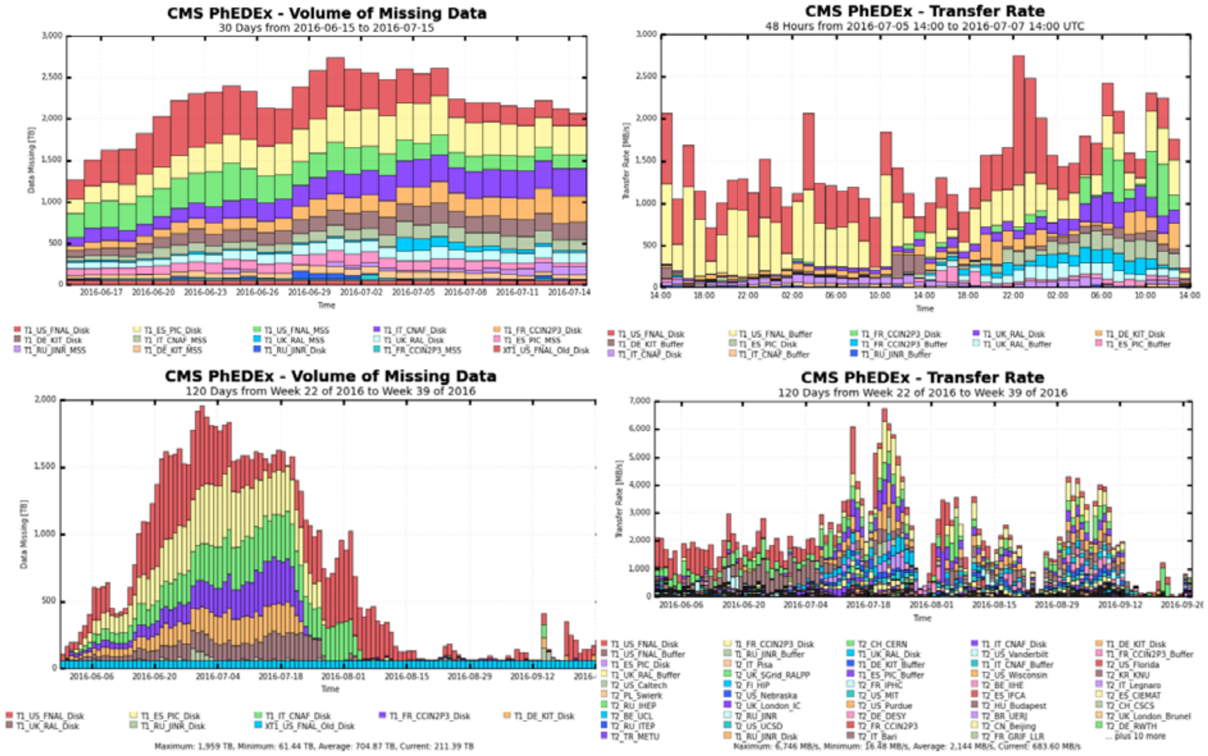

Figure 4. Raw data backlog issue revealed the importance of the clockwork (co)operation among the different systems involved in distributed computing. The two plots on top are showing the data backlog accumulated before the intervention (left) and the exporting rates via the gridftp doors (right) for CMS. Bottom plots shows the impact of all the actions performed boosting the data export rate (right) and the final consolidation of the data to the destination sites (left)

has the ability to access at the same time on the same node the big data repositories of the experiments, the experiment software mounted via CVMFS[6] and their personal areas as this three components are available via FUSE (Filesystem in User Space) in a file-system-like fashion, with the advantage of the sync and share capabilities via the personal storage area on CERNBOX. This means that collaborative frameworks can be build among research groups to have immediate access by several people to the results coming from the analysis farms and hence work in a common and completely shareable fashion.

6. Facilitating Science: linking HTC, HPC and Distributed Computing

One of the key points of storage solutions is to facilitate science and provide the means to support the workflows needed by the scientific communities. A success case and a good example of the efforts of the computing community and scientific community getting results together has been achieved with the COMPASS experiment at CERN. COMPASS had granted a project to make use of the BlueWaters National Centre for Supercomputing Applications in Chicago (IL) . They needed to reprocess experiment data during a specific period of time. These data was stored on tape and disk on CASTOR and also on disk on EOS. They were making this copy manually via standard put commands via interactive commands but this was demonstrated not to be time nor cost effective as it was purely manual and involved many

[6] CVMFS stands for CernVM File System and it is a scalable, reliable and low-maintenance software distribution service via squid servers

parties: pre-staging from tape before transferring, authentication and authorisation, no protection against failures, no data integrity checks etc. We proposed a solution which we have been using since years in WLCG and is to interface CASTOR and EOS endpoints with remote storage endpoints. Blue Waters storage system was connected with CERN Storage Systems via FTS3. This was implemented rapidly as globus toolkit is available in most of the Supercomputing centres and big data volumes were transferred from CERN to Blue Waters with all the benefits of FTS3: bulk requests, throughput scaling mechanisms, credential delegation, auto-retries in case of failures or corruptions, etc. This implementation opens a door for HPC environments to link with our HTC and Distributed Computing tools and expertise.

7. Summary: Easy data access. Simplify analysis. Facilitate science

Data is becoming big, in all aspects, for all experiments and all the use cases. Data is precious as sometimes cannot be derived. Data is the existing reason for storage. Data is meant to be accessed and read to provide scientists the blocks for the analysis, investigations and discoveries. Our mission is to provide easy means to access data and provide the requirements and performance the experiments and users ask for. These requirements are not always common and we should make our systems flexible enough to cope or adopt alternate solutions if necessary. Our storage systems successfully catered with LHC and non-LHC experiment needs during the last years. We continued evolving them during the LS1 to successfully face the upcoming Run2 challenge. Our mission and visions are still intact: provide Large Scale Storage for physics, reliable platforms for services and tools to facilitate user analysis for the scientific community.

References

[1] Lo Presti G et al CASTOR A Distributed Storage Resource Facility for High Performance Data Processing at CERN *IEEE Conference on Mass Storage Systems and Technologies* 2007

[2] Peters A and Janyst L Exabyte scale storage at CERN *J. Phys. Conf. Ser.* **331** 2011

[3] ROOT an Object-Oriented Data Analysis Framework *http://root.cern.ch*

[4] Weil S et al Ceph: a scalable, high-performance distributed file system *Proceedings of the 7th Conference on Operating Systems Design and Implementation* 2006

[5] ATLAS Collaboration PanDA: Exascale Federation of Resources for the ATLAS Experiment at the LHC *EPJ Web Conf.* **108** 2016

[6] Espinal X et al Disk storage at CERN: Handling LHC data and beyond *J. Phys. Conf. Ser.* **513** 2014

[7] Peters A, Sindrilaru E and Adde G EOS as the present and future solution for data storage at CERN *J. Phys. Conf. Ser.* **664** 2015

[8] Mascetti L, Gonzalez H, Lamanna M, Moscicki J and Peters A CERNBox and EOS: end-user storage for science *J. Phys. Conf. Ser.* **664** 2015

[9] Egeland R, Wildish T and Huang C PhEDEx data service *J. Phys. Conf. Ser.* **219** 2010

[10] Kiryanov A, Ayllon A, Salichos M and Keeble O FTS3 A File Transfer Service for Grids, HPCs and Clouds *PoS ISGC* 2015